Numerical Integration of Systems of Stiff Nonlinear Differential Equations

By I. W. SANDBERG and H. SHICHMAN

(Manuscript received November 29, 1967)

In connection with the design of transistor circuits, for example, it is frequently necessary to obtain a numerical solution of a system of nonlinear ordinary differential equations. In some cases, these equations possess a property that leads to intolerable computational requirements relative to the use of standard predictor-corrector techniques or general linear multipoint formulas of open type.

Here we describe an alternative approach which has been used to solve some practical problems by permitting dramatic step-size increases (for example, a factor of 10^4). The approach is developed in a way which provides some detailed understanding of why it is useful.

I. INTRODUCTION

In connection with the design of transistor circuits, for example, it is often necessary to obtain a numerical solution of a system of nonlinear differential equations

$$\dot{x} + f(x, t) = 0, \qquad t \geqq 0, \qquad [x(0) = x_0] \tag{1}$$

in which x and $f(x, \cdot)$ are N-vector-valued functions of t. The simplest numerical-integration formula which can be in principle used for this purpose is Euler's formula:

$$y_{n+1} = y_n + hy'_n, \qquad n \geqq 0 \tag{2}$$

in which h, a positive number, is the step size; $y_0 = x_0$;

$$y'_n = -f(y_n, nh) \quad \text{for} \quad n \geqq 0;$$

and y_n is of course the approximation to $x(nh)$ for $n \geqq 1$.

It is frequently the case that $f(x, \cdot)$ possess a property that leads to computational requirements consistent with the use of (2) that are intolerable. To see clearly how this situation can arise suppose that

the solution of (1) is desired over some finite interval $[0, \tau]$, and consider the very special case in which $f(x, t) = Ax$ with A an $N \times N$ matrix possessing distinct eigenvalues $\{a_i\}$ all of which have positive real parts. Then using the fact there exists a nonsingular transformation T such that

$$A = TDT^{-1}, \qquad D = \text{diag}\,(a_1, a_2, \cdots, a_N) \tag{3}$$

we have

$$y_{n+1} = T(1_N - hD)T^{-1}y_n, \qquad n \geqq 0, \qquad [y_0 = x_0] \tag{4}$$

in which 1_N is the identity matrix of order N. From (4)

$$y_k = T(1_N - hD)^k T^{-1}x_0, \qquad k \geqq 0. \tag{5}$$

Since

$$x(kh) = Te^{-Dkh}T^{-1}x_0, \qquad k \geqq 0 \tag{6}$$

it is evident that the numerical solution is "acceptable" if h is so small that $(1 - ha_i)^k$ is an "acceptable" approximation to $e^{-a_i kh}$ for all i and all values of k for which $0 \leqq kh \leqq \tau$. On the other hand if for some value of i

$$|\, 1 - ha_i \,| = 1, \quad \text{or} \quad |\, 1 - ha_i \,| > 1$$

then for at least one initial condition vector x_0, $\{||\, y_k \,||\}_0^\infty$ ($||\cdot||$ denotes the usual Euclidian norm) does not approach zero as $k \to \infty$ or is unbounded, respectively [that is, (2) is numerically unstable]. Therefore if the sequence $\{y_k\}$ defined by (4) is to be a good approximation to the samples of the solution of (1) with $f(x, t) = Ax$, it is certainly necessary that

$$|\, 1 - ha_i \,| < 1 \quad \text{for all } i. \tag{7}$$

Moreover, in order to fully determine the character of the solution of the differential equation, it is reasonable to assume that τ, the length of the interval over which the solution is desired, is proportional (by some factor c such as 3 or 10) to the reciprocal of $\min_i \text{Re}(a_i)$ (that is, proportional to the largest time constant of the system). Thus in addition to (7) we have

$$\tau = c[\min_i \text{Re}\,(a_i)]^{-1}. \tag{8}$$

A lower bound on the number of evaluations of (2) necessary to compute the solution is τ/h where h satisfies (7). If all of the a_i are

real, the smallest lower bound is simply

$$\tfrac{1}{2}c \frac{\max_i (a_i)}{\min_i (a_i)}. \tag{9}$$

It is a simple matter to give examples of, for instance, positive-element linear RC networks governed by a state equation of the form $\dot{x} + Ax = 0$ for which the bound (9) can be made arbitrarily large by choosing the value of one capacitor to be arbitrarily small. Tbus, from tbe practical viewpoint, computation based on (2) can be impossible as a result of tbe presence of parasitic circuit elements tbat bave no really significant effect on the circuit performance! It is not surprising tberefore that a more complex and pressing problem of the same type arises in connection with the numerical solution of the nonlinear differential equations of transistor circuits, as a result of, for example, tbe parasitic capacitors associated with the models of transistors. For many practical circuits of this type, computation time estimates, based upon use of (2) and a modern high-speed computer, are about 1000 hours.

The well-known basic problem described above arises not only in connection of the use of (2), but (as can easily be shown) is encountered also in attempts to use more general integration formulas of open type[1, 2]

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=0}^{p} b_k y'_{n-k}, \tag{10}$$

or predictor-corrector techniques[1, 2] such as

$$\begin{aligned} y_{n+1}{}^{(p)} &= y_{n-1} + 2hy'_n \\ y_{n+1}{}^{(c)} &= y_n + \tfrac{1}{2}h(y'_n + y'_{n+1}{}^{(p)}). \end{aligned} \tag{11}$$

The fundamental difficulty associated with the integration of "stiff equations" results from the restrictions that must be imposed on h in order to insure numerical stability.

The purpose of this paper is to consider the properties of alternative numerical methods for obtaining solutions of equations of the form (1). Our principal objective is to present some analytical results that shed some light on the properties of a class of numerical-integration techniques that have been used to solve practical transistor circuit problems by permitting dramatic step-size increases (for example, a factor of 10^4) relative to the methods defined by (10) and (11).

More explicitly, attention is focused on "large-h algorithms" based

on, or derived from, the standard formula of closed type

$$y_{n+1} = y_n + hy'_{n+1} \tag{12}$$

which is a special case of the general multipoint formula

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + h \sum_{k=-1}^{p} b_k y'_{n-k} \tag{13}$$

with $b_{-1} \neq 0$. There is an extensive body of information concerning (12) in the numerical-analysis literature only for the case in which h is "sufficiently small."

II. INTEGRATION FORMULA

If we use the numerical-integration formula

$$y_{n+1} = y_n + hy'_{n+1} \tag{12}$$

in an attempt to compute the solution of (1), then y_{n+1} is defined implicitly in terms of y_n through

$$y_{n+1} + hf[y_{n+1}, (n+1)h] = y_n, \qquad n \geq 0, \qquad [y_0 = x_0]. \tag{13}$$

For the special case considered in Section I, in which $f(x, t) = Ax$ and $A = TDT^{-1}$, we have

$$y_{n+1} = T(1_N + hD)^{-1}T^{-1}y_n, \qquad n \geq 0 \tag{14}$$

and to the extent that $(1 + ha_i)^{-1}$ is a good approximation to $e^{-a_i h}$, (13) generates an acceptable numerical solution of the differential equation. More explicitly (13) generates the *exact* solution of the differential equation

$$\dot{x} + \tilde{A}x = 0, \qquad t \geq 0, \qquad [x(0) = x_0] \tag{15}$$

in which $\tilde{A} = T\tilde{D}T^{-1}$ and $e^{-\tilde{D}h} = (1_N + hD)^{-1}$.

Let us suppose now that all of the a_i are real and that ha_i is very small relative to unity for i belonging to a proper subset $\mathcal{S}$ of $\mathfrak{N} \triangleq \{1, 2, \cdots, N\}$, and that ha_i is very large relative to unity for i belonging to the complement $\bar{\mathcal{S}}$ of $\mathcal{S}$ with respect to $\mathfrak{N}$. Then for all $i \in \mathcal{S}$, $\tilde{a}_i$, the ith element of $\tilde{D}$ is very nearly a_i, while for all $i \in \bar{\mathcal{S}}$, $\tilde{a}_i < a_i$ and $\tilde{a}_i$ is very much larger than all of the $\tilde{a}_i$ for which $i \in \mathcal{S}$.

In other words, roughly speaking, (13) generates a solution to a differential equation governing a system similar to that governed by $\dot{x} + Ax = 0$; the former system has virtually the same low-frequency performance and less pronounced high-frequency performance. To

look at the situation in still another way, in using (13) we are able to (*i*) break away from an extremely restrictive requirement on h for numerical stability, such as (7), and (*ii*) trade step-size for accuracy of high-frequency solution components.

The simple heuristic argument given above suggests that the use of (12) can lead to a considerable increase in permissible step sizes for a class of nonlinear transistor circuit problems in which typically the Jacobian matrix $\partial f(x, t)/\partial x$ of $f(x, t)$ along the solution of (1) possesses only real eigenvalues which are widely separated. This argument is supported by a proposition, proved in Section IV, which is concerned with the case in which there exists a constant $m > 0$ such that (with $\langle \cdot , \cdot \rangle$ denoting the usual inner product)

$$\langle y, f(y, nh) - f(0, nh) \rangle \geqq m \| y \|^2 \tag{16}$$

for all $n \geqq 0$ and all y. If this condition is satisfied for all $h > 0$, which for the scalar case is true if

$$\frac{\partial f(y, t)}{\partial y} \geqq m$$

for all t and all y, if $\| f(0, t) \| \to 0$ as $t \to \infty$ or if $\| f(0, t) \|$ is uniformly bounded on $[0, \infty)$, then (as can easily be shown) $\| x(t) \| \to 0$ as $t \to \infty$ or $\| x(t) \|$ is uniformly bounded on $[0, \infty)$, respectively. The Proposition asserts that if (16) is met and y_{n+1} is defined for $n \geqq 0$ by (13) ,then

$$\| y_n \| \leqq (1 + mh)^{-n} \| x_0 \| + \sum_{k=0}^{n-1} (1 + mh)^{-(k+1)} \| hf[0, (n - k)h] \|$$

for all $n \geqq 1$, which implies that (13) is numerically stable for all h in the sense that for all h, $\| f(0, nh) \| \to 0$ as $n \to \infty$ implies that $y_n \to 0$ as $n \to \infty$ and $\{\| f(0, nh) \|\}_0^\infty$ bounded implies that $\{y_n\}_0^\infty$ is bounded.

Although the result stated above does not provide quantitative information concerning the errors incurred in using (13), it does show under a reasonable assumption concerning $f(x,t)$ that unlike all formulas (10) of open type and unlike predictor-corrector methods such as (11), (13) defines for any step size a sequence $\{y_n\}$ which is consistent with either or both of two possible basic properties of the true solution.

The discussion above does not take into account the fact that at each step errors are inevitably introduced in solving the equation

$$y_{n+1} + hf[y_{n+1}, (n + 1)h] = y_n \tag{17}$$

for y_{n+1}. Consider the result of using the iteration scheme

$$y_{n+1}^{(k+1)} = y_n - hf[y_{n+1}^{(k)}, (n+1)h], \qquad y_{n+1}^{(0)} = y_n$$

which is the usual method described [1, 2] in connection with the theory of integration formulas of closed type. For the linear case [that is, for $f(x, t) = Ax$],

$$\begin{aligned} y_{n+1}^{(k)} &= \sum_{i=0}^{k} (-hA)^i y_n \\ &= T \sum_{i=0}^{k} (-hD)^i T^{-1} y_n . \end{aligned} \tag{18}$$

Therefore, if $\tilde{y}_1$ denotes the approximation to y_1 computed from y_0 after k_1 iterations, and if $\tilde{y}_2$ denotes the approximation to y_2 computed from $\tilde{y}_1$ after k_2 iterations and so forth, then

$$\tilde{y}_K = T\Theta_{k_K}\Theta_{k_{K-1}} \cdots \Theta_{k_2}\Theta_{k_1}T^{-1}y_0$$

in which

$$\Theta_{k_p} = \operatorname{diag}\left(\sum_{i=0}^{k_p} (-ha_1)^i, \cdots, \sum_{i=0}^{k_p} (-ha_N)^i\right).$$

Since (assuming now that all of the a_i are real)

$$\left| \sum_{i=0}^{k_p} (-ha_i)^i \right| > 1 \tag{19}$$

provided that $ha_i > 2$ and $k_p \geqq 1$, if $ha_i > 2$ for some i, then $\| \tilde{y}_k \| \to \infty$ as $k \to \infty$ for some initial condition y_0, independent of the sequence $k_1, k_2, \cdots$. Therefore the usual iteration method will reintroduce the numerical instability for insufficiently small h which it is our objective to avoid.*

Let us consider now a different and more general approach of solving (17) for y_{n+1}. Assume that there exists a positive constant l such that $f(y, nh)$ satisfies the Lipshitz condition

$$\| f(y_1, nh) - f(y_2, nh) \| \leqq l \| y_1 - y_2 \|$$

for all $n \geqq 0$ and all y_1 and y_2. Suppose also that the smallest eigenvalue of the symmetric part of the Jacobian matrix $\partial f(y, nh)/\partial y$ of

* Similar instability results for the nonlinear case can be proved. But since this is hardly surprising, we shall not consider the matter further.

$f(y, nh)$ is bounded from below hy m, a positive constant, for all y and all n.

Ideally, we would like to determine the sequence $\{y_n\}_0^\infty$ defined by

$$y_{n+1} + hf[y_{n+1}, (n+1)h] = y_n, \qquad n \geqq 0.$$

Suppose that we determine instead a sequenee $\{\tilde{y}_n\}_0^\infty$ such that $\tilde{y}_0 = y_0$

$$\| \tilde{y}_n - y_n^* \| \leqq \epsilon$$

and

$$y_{n+1}^* + hf[y_{n+1}^*, (n+1)h] = \tilde{y}_n$$

for $n \geq 0$ in which ϵ is an arbitrary positive constant independent of n. In other words, suppose that at each step the *local* error in solving for y_{n+1} is at most ϵ. Then, according to Theorem 1 (Section IV)

$$\| \tilde{y}_n - y_n \| \leqq \epsilon(1 + hl)(1 + hm)^{-1} \sum_{k=0}^{n-1} (1 + hm)^{-k}$$

for all $n \geq 1$, which of course implies the uniform bound

$$\| \tilde{y}_n - y_n \| \leqq \epsilon(1 + hl)(hm)^{-1}, \qquad n \geqq 1. \tag{20}$$

Our assumption concerning $\partial f(y,nh)/\partial y$ implies that the condition

$$\langle y, f(y, nh) - f(0, nh) \rangle \geqq m \| y \|^2$$

of the Proposition is met. Thus it follows from the Proposition and (20) that if the local error in solving for y_{n+1} is held to within ϵ at eaeh step, then the algorithm is numerically stable for all h in the sense that for all h (*i*) $\{\| f(0, nh) \|\}_0^\infty$ hounded implies that $\{\tilde{y}_n\}_0^\infty$ is bounded, and (*ii*) $\| f(0, nh \| \to 0$ as $n \to \infty$ implies that for any $\delta > 0$ there exists an n_0 such that $\| \tilde{y}_n \| \leqq \epsilon(1 + hl)(hm)^{-1} + \delta$ for all $n \geqq n_0$.

The combination of this stability result and the heuristic argument of Section I strongly suggests that the following approach should permit the use of considerably increased step sizes with acceptable accuracy, for many of the "widely-separated eigenvalue" problems described earlier. Referring to (17), solve for y_{n+1} at each step using, say, the Newton-Raphson technique;* iterate until some norm of

* After the work reported here had heen completed, A. N. Willson, Jr. brought to our attention a preprint of a paper hy R. Willoughby and several of his colleagues at IBM, in which an approach of this type is suggested. The preprint does not contain the principal results of this paper, the material of Section IV.

the difference between the last two iterates is not greater than some small preseribed constant.

In partieular, notiee that for $f(x, t) = Ax$, this approach, using the Newton-Raphson iteration procedure, reduces to the use of the formula $y_{n+1} = (1_N + hA)^{-1}y_n$ (that is, to equation 14).

The teehnique described above has provided a significant reduction in total eomputation time for several types of praetieal prohlems. It was used, for example, to solve the system of differential equations governing the eireuit of Fig. 1, an oseillator designed to supply a 1 ke signal. The 16 G Western Electrie 100 Mc. silicon transistor of Fig. 1 was represented by a charge-control model (see Section 6.2, pp. 556–557 of Koehler[3]) using two nonlinear eharge-controlled voltage sources, with the result that the system of equations for the eireuit is of order 5.

Motivated by the fact that the local-truneation error for formula (12) is $\frac{1}{2}h^2\ddot{x}(\xi)$ for some $\xi \in [nh, (n + 1)h]$, the following method was used (for this problem as well as for others) to eontrol the step size. Let e denote the largest of the magnitudes of the elements of the veetor of seeond differenees associated with the most recently eomputed point. If $e \in [\frac{1}{4}\bar{e}, \bar{e}]$ (for this prohlem $\bar{e}$ was taken to be 10^{-4}), then the point is aeeepted; if $e > \bar{e}$, then the point is rejeeted and the calculation is repeated with h replaeed with $\frac{1}{2}h$. If $e < \frac{1}{4}\bar{e}$, then the point is aeeepted and h is replaeed by $2h$ in the computation of the next point. Average step-size inereases of ahout 10^4 (relative to, for example, the use of a forth-order predietor-correetor method) were ohtained for this problem (see Fig. 2).

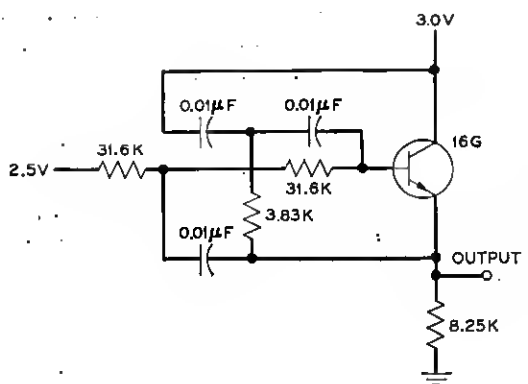

Fig. 1 — One-kilocycle oscillator using a 16G "100 megacycle" transistor.

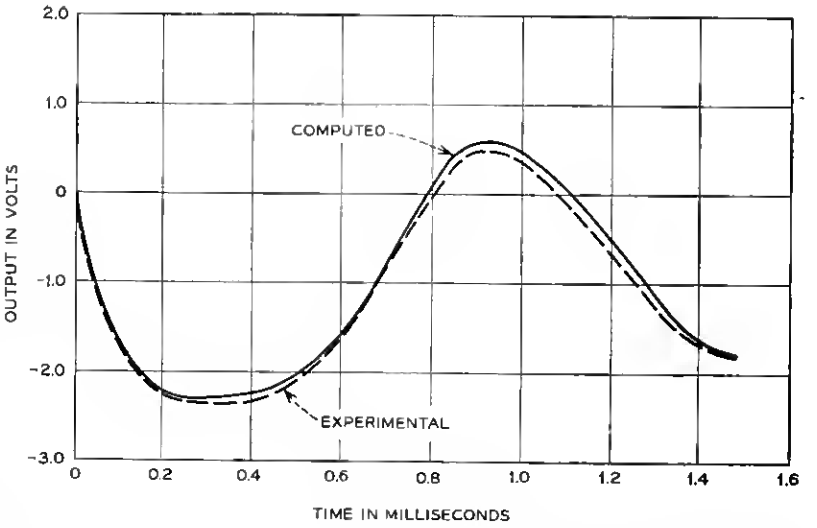

Fig. 2 — Comparison of computed and experimental response of the oscillator shown in Fig. 1.

III. AN EXPLICIT INTEGRATION FORMULA

Of particular interest in connection with the approach described above is the numerical-integration formula

$$y_{n+1} = y_n - \{1_N + hf'[y_n, (n+1)h]\}^{-1} hf[y_n, (n+1)h], \qquad n \geqq 0, \quad [y_0 = x_0] \tag{21}$$

which is obtained from

$$Y_{n+1} + hf[Y_{n+1}, (n+1)h] = Y_n \tag{22}$$

by replacing Y_n by y_n and using as the approximation y_{n+1} to Y_{n+1} the result obtained by using one step of the Newton-Raphson iteration scheme with y_n the initial point. That is, with

$$\begin{aligned} Q(z) &= z + hf[z, (n+1)h] - y_n, \\ y_{n+1} &= y_n - [Q'(z)|_{z=y_n}]^{-1} Q(z)|_{z=y_n}. \end{aligned} \tag{23}$$

In spite of its relative simplicity, it has been found that formula (21) is useful for solving problems of the type that we have been considering. For the problem of Fig. 1, it has led to an average step size increase of about 10^3.

In view of the simplicity of formula (21), and especially in view of the fact that y_{n+1} is defined explicitly in terms of y_n, it deserves special consideration.

Theorem 2 (Section IV) asserts that for any $h > 0$ there exist positive constants k_1 and k_2 such that $k_1 < 1$ and

$$|| y_n || \leqq k_1^n || y_0 || + hk_2 \sum_{k=0}^{n-1} k_1^k || f[0, (n-k)h] || \tag{24}$$

for $n \geqq 1$, provided that the Jacobian matrix $\partial f(y, nh)/\partial y$ satisfies certain conditions. For the scalar case, these conditions reduce to:

(*i*) there exist positive constants k and m such that

$$m \leqq \frac{\partial f(y, nh)}{\partial y} \leqq k$$

for all y and all $n \geq 1$

(*ii*)
$$2\frac{\partial f(y, nh)}{\partial y} - \frac{\partial f(\alpha y, nh)}{\partial y} \geqq 0$$

for all y, $n \geqq 1$ and $\alpha \in [0, 1]$.

Clearly, under these conditions, $y_n \to 0$ as $n \to \infty$ if $f(0, nh) \to 0$ as $n \to \infty$ and $\{y_n\}$ is bounded if $\{||f(0, nh)||\}$ is bounded.

The function $f(y, nh)$ of Fig. 3 is one for which conditions (*i*) and (*ii*) are clearly met. If condition (*ii*) is not met, then (24) need not follow. To show this, consider, for example, the function of Fig. 4

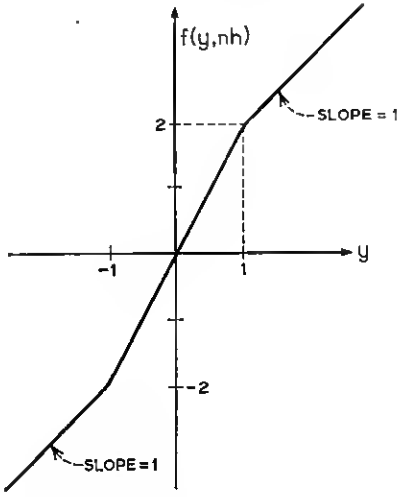

Fig. 3 — Definition of $f(y, nh)$ for all n.

which meets condition (i), but not condition (ii). We have from (21):

$$h = 1 \quad \text{and} \quad y_0 = 1 \quad \text{imply that} \quad y_1 = -1$$

and

$$h = 1 \quad \text{and} \quad y_1 = -1 \quad \text{imply that} \quad y_2 = 1$$

from which it is clear that for this function $y_n = (-1)^n$ if $h = 1$ and $y_0 = 1$, which of course implies [here $f(0, nh) = 0$ for all n] that (24) is not satisfied. Thus we see that if condition (ii) is not met, then (24) need not follow.

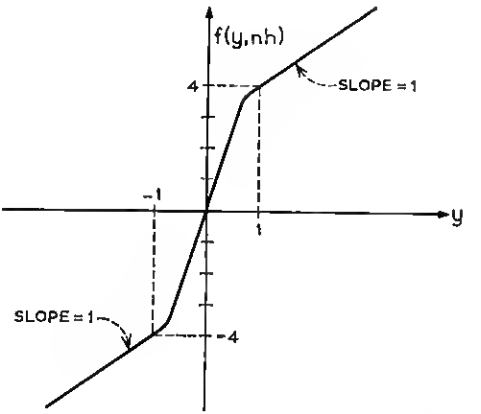

Fig. 4 — Alternate definition of $f(y, nh)$ for all n.

IV. PROPOSITION AND THEOREMS*

Proposition: If $\{y_n\}$ satisfies

$$y_{n+1} + hf[y_{n+1}, (n + 1)h] = y_n, \qquad n \geqq 0$$

and if there exists an $m > 0$ such that

$$\langle y, f(y, nh) - f(0, nh) \rangle \geqq m \| y \|^2, \qquad n \geqq 0$$

for all real y, then

$$\| y_n \| \leqq (1 + mh)^{-n} \| y_0 \| + \sum_{k=0}^{n-1} (1 + mh)^{-k} \| hf[0, (n - k)h] \|$$

for $n \geq 1$.

* Throughout this section, $\|\cdot\|$ denotes the usual Euclidean Norm and $\langle \cdot, \cdot \rangle$ denotes the corresponding usual scalar product.

Proof: Clearly,

$$\begin{aligned}\langle y_{n+1} , y_n - hf[0, (n+1)h]\rangle &= \langle y_{n+1} , y_{n+1} + hf[y_{n+1} , (n+1)h] - hf[0, (n+1)h]\rangle \\ &\geqq (1 + mh) \| y_{n+1} \|^2,\end{aligned}$$

and, by the Schwarz inequality,

$$\langle y_{n+1} , y_n - hf[0, (n+1)h]\rangle \leqq \| y_{n+1} \| \cdot \| y_n \| + \| y_{n+1} \| \cdot \| hf[0, (n+1)h] \|.$$

Thus

$$\| y_{n+1} \| \leqq (1 + mh)^{-1} \| y_n \| + (1 + mh)^{-1} \| hf[0, (n+1)h] \|$$

from which we have

$$\| y_n \| \leqq (1 + mh)^{-n} \| y_0 \| + \sum_{k=0}^{n-1} (1 + mh)^{-(k+1)} \| hf[0, (n-k)h] \|$$

for $n \geqq 1$, which completes the proof.

Definition: Let $\lambda(y, nh)$ denote the smallest eigenvalue of the symmetric part of $\partial f(y, nh)/\partial y$.

Theorem 1: Suppose that there exists a constant m such that $\lambda(y, nh) \geqq m > 0$ for all $n \geqq 0$ and all y, and that there exists a constant l such that

$$\| f(y_1 , nh) - f(y_2 , nh) \| \leqq l \| y_1 - y_2 \|$$

for all $n \geqq 0$ and all y_1 and y_2. If $\{y_n\}$ satisfies

$$y_{n+1} + hf[y_{n+1} , (n+1)h] = y_n , \qquad n \geqq 0$$

if, with ϵ a positive constant, $\{\tilde{y}_n\}$ satisfies

$$\| \tilde{y}_n - y_n^* \| \leqq \epsilon \quad \text{for} \quad n \geqq 0 \quad \text{with}$$
$$y_{n+1}^* + hf(y_{n+1}^* , (n+1)h) = \tilde{y}_n$$

then

$$\| \tilde{y}_n - y_n \| \leqq (1 + hm)^{-n} \| \tilde{y}_0 - y_0 \| + (1 + hm)^{-1}(1 + hl) \epsilon \sum_{k=0}^{n-1} (1 + hm)^{-k}$$

for $n \geqq 1$.

Proof: We have for $n \geqq 0$:

$$\tilde{y}_{n+1} + hf[\tilde{y}_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h] = \tilde{y}_n + (\tilde{y}_{n+1} - y^*_{n+1})$$

and

$$y_{n+1} + hf[y_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h]$$
$$= y_n + hf[y_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h] - hf[y_{n+1}, (n+1)h].$$

Therefore

$$\begin{aligned}\tilde{y}_{n+1} - y_{n+1} &+ hf[\tilde{y}_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h] \\ &- hf[y_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h] \\ &= \tilde{y}_n - y_n + (\tilde{y}_{n+1} - y^*_{n+1}) + hf(y_{n+1}, (n+1)h) \\ &- hf[y_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h]. \qquad (25)\end{aligned}$$

With f'_s the symmetric part of $\partial f(y, nh)/\partial y$, the inner-product of $(\tilde{y}_{n+1} - y_{n+1})$ with the left side of (25) is

$$\| \tilde{y}_{n+1} - y_{n+1} \|^2 + h\Big\langle \tilde{y}_{n+1} - y_{n+1}, \int_0^1 f'_s\{\alpha[\tilde{y}_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1})]$$
$$+ (1-\alpha)[y_{n+1} + (y^*_{n+1} - y_{n+1})], (n+1)h\}\, d\alpha(\tilde{y}_{n+1} - y_{n+1})\Big\rangle, \qquad (26)$$

since

$$f[\tilde{y}_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h] - f[y_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h]$$
$$= \int_0^1 \frac{\partial f[y, (n+1)h]}{\partial y}\bigg|_{y=\alpha[\ ]+(1-\alpha)[\ ]} d\alpha(\tilde{y}_{n+1} - y_{n+1}).$$

Expression (26) is bounded from below by

$$(1 + hm) \| \tilde{y}_{n+1} - y_{n+1} \|^2.$$

By the Schwarz inequality, the inner-product of $(\tilde{y}_{n+1} - y_{n+1})$ with the right side of (25) is bounded from above by

$$\| \tilde{y}_{n+1} - y_{n+1} \| \cdot \| \tilde{y}_n - y_n \|$$
$$+ \| \tilde{y}_{n+1} - y_{n+1} \| \cdot \| \tilde{y}_{n+1} - y^*_{n+1} \| + \| \tilde{y}_{n+1} - y_{n+1} \|$$
$$\cdot \| hf[y_{n+1}, (n+1)h] - hf[y_{n+1} + (y^*_{n+1} - \tilde{y}_{n+1}), (n+1)h] \|,$$

which is bounded from above by

$$\| \tilde{y}_{n+1} - y_{n+1} \| \cdot \| \tilde{y}_n - y_n \| + \| \tilde{y}_{n+1} - y_{n+1} \| (\epsilon + hl\epsilon).$$

Thus,

$$\| \tilde{y}_{n+1} - y_{n+1} \| \leqq (1 + hm)^{-1} \| \tilde{y}_n - y_n \| + (1 + hm)^{-1}(1 + hl)\epsilon,$$

from which it follows that

$$\| \tilde{y}_n - y_n \| \leqq (1 + hm)^{-n} \| \tilde{y}_0 - y_0 \| + (1 + hm)^{-1}(1 + hl)\, \epsilon \sum_{k=0}^{n-1} (1 + hm)^{-k}$$

for all $n \geqq 1$.

Theorem 2: If $\{y_n\}$ satisfies

$$y_{n+1} = y_n - \{1_N + hf'[y_n, (n+1)h]\}^{-1} hf[y_n, (n+1)h]$$

for $n \geqq 0$, *if*

(*i*) *there exists a constant* $k < \infty$ *such that*

$$\left\| \frac{\partial f(y, nh)}{\partial y} \right\| \leqq k$$

for all $n \geqq 1$ *and all* y

(*ii*) *there exists a constant* $m > 0$ *such that* $\lambda(y, nh) \geqq m$ *for all* $n \geqq 1$ *and all* y

(*iii*) *with* $F \triangleq hf'[y, (n+1)h]$ *and* $F_\alpha \triangleq hf'[\alpha y, (n+1)h]$, *the symmetric part of* $\{(2F - F_\alpha)F_\alpha^t\}$ *is* nonnegative definite for all* y, *all* n, *and all* $\alpha \in [0, 1]$,

then there exist positive constants k_1 *and* k_2 *such that* $k_1 < 1$ *and*

$$\| y_n \| \leqq k_1^n \| y_0 \| + hk_2 \sum_{k=0}^{n-1} k_1^k \| f[0, (n-k)h] \| \quad \text{for all} \quad n \geqq 1.$$

Proof: We have

$$y_{n+1} = y_n - \{1_N + hf'[y_n, (n+1)h]\}^{-1}\{hf[y_n, (n+1)h] - hf[0, (n+1)h]\} - \{1_N + hf'[y_n, (n+1)h]\}^{-1} hf[0, (n+1)h];$$

hence

$$\| y_{n+1} \| \leqq \left\| 1_N - (1_N + F)^{-1} \int_0^1 F_\alpha \, d\alpha \right\| \cdot \| y_n \| + \| (1_N + F)^{-1} \| \cdot \| hf[0, (n+1)h] \| \qquad (27)$$

* The superscript t denotes matrix transposition.

with the understanding that F and F_α are evaluated at $y = y_n$, since

$$hf[y_n\ ,\ (n+1)h] - hf[0,\ (n+1)h] = \int_0^1 hf'[\alpha y_n\ ,\ (n+1)h]\ d\alpha\ y_n\ .$$

We now prove that there exists $k_1 \in (0, 1)$ such that

$$\left\| 1_N - (1_N + F)^{-1} \int_0^1 F_\alpha\ d\alpha \right\| \leqq k_1$$

for all n and all y_n.

From condition (*iii*), with V an arbitrary N-vector,

$$\langle (2F^t - F_\alpha^t)V,\ F_\alpha^t V\rangle \geqq 0$$

or

$$\langle 2F^t V,\ F_\alpha^t V\rangle - \langle F_\alpha^t V,\ F_\alpha^t V\rangle \geqq 0$$

which implies that

$$\| F_\alpha^t V \|^2 - 2\langle F^t V,\ F_\alpha^t V\rangle + \| F^t V \|^2 \leqq \| F^t V \|^2$$

or

$$\| (F^t - F_\alpha^t)V \|^2 \leqq \| F^t V \|^2.$$

In view of conditions (*i*) and (*ii*), it is evident that there exists a $\xi \in (0, 1)$ such that

$$-2\langle F_\alpha^t V,\ V\rangle \leqq -(1-\xi)\| V \|^2 - 2(1-\xi)\langle F^t V,\ V\rangle - (1-\xi)\| F^t V \|^2$$

for all α, n, y_n, and V. Therefore

$$\| (F^t - F_\alpha^t)V \|^2 - \|F^t V \|^2 - 2\langle F_\alpha^t V,\ V\rangle$$
$$\leqq -(1-\xi)\| V \|^2 - 2(1-\xi)\langle F^t V,\ V\rangle - (1-\xi)\| F^t V \|^2$$

which is the same as

$$\| V \|^2 + \| (F^t - F_\alpha^t)V \|^2 + 2\langle (F^t - F_\alpha^t)V,\ V\rangle$$
$$\leqq \xi \| V \|^2 + 2\xi\langle F^t V,\ V\rangle + \xi \| F^t V \|^2$$

or

$$\| (1_N + F^t - F_\alpha^t)V \|^2 \leqq \xi \| (1_N + F^t)V \|^2.$$

With $U = (1_N + F^t)V$, we have

$$\| (1_N + F^t - F_\alpha^t)(1_N + F^t)^{-1}U \|^2 \leqq \xi \| U \|^2. \tag{28}$$

Since (28) is satisfied for all U,

$$|| (1_N + F^t - F_\alpha^t)(1_N + F^t)^{-1} || \leqq k_1$$

with $k_1 = (\xi)^{\frac{1}{2}}$. However,

$$|| (1_N + F^t - F_\alpha^t)(1_N + F^t)^{-1} || = || (1_N + F)^{-1}(1_N + F - F_\alpha)||,$$

and

$$\left|\left| 1_N - (1_N + F)^{-1} \int_0^1 F_\alpha \, d\alpha \right|\right|$$

$$\leqq \int_0^1 || (1_N + F)^{-1}(1_N + F - F_\alpha) || \, d\alpha \leqq k_1 .$$

Consider now $||(1_N+F^{-1}||$. Since for any V

$$|| (1_N + F)V ||^2 = || V ||^2 + 2\langle FV, V\rangle + || FV ||^2 \geqq (1 + 2hm) || V ||^2,$$

it follows at once that

$$|| (1_N + F)^{-1} || \leqq (1 + 2hm)^{-\frac{1}{2}}.$$

Thus with $k_2 = (1+2hm)^{-\frac{1}{2}}$

$$|| y_{n+1} || \leqq k_1 || y_n || + k_2 || hf[0, (n + 1)h] ||$$

from which we obtain the bound on $||y_n||$ stated in the theorem.

V. FINAL REMARKS

The algorithm described in this paper is a marriage of two standard techniques, the use of a well-known closed-type numerical-integration formula and the Newton-Raphson iteration procedure. It is clear that the approach is of use in connection with a certain class of practical problems, and, what is of at least as much importance, we have some detailed understanding of why the algorithm is useful.

It is also clear that some natural generalizations and extensions of the approach, such as using different closed-type formulas* or different methods of solving systems of nonlinear equations, will lead to more efficient techniques. Finally, since there are several alternate approaches available which are also of use in certain cases (see Pope,

* For example, for the trapezoidal rule $y_{n+1} = y_n + \frac{1}{2}h(y_n' + y_{n+1}')$ and for $f(x, t) = Ax$, we have $y_{n+1} = T\Xi T^{-1}y_n$, in which $\Xi = \text{diag}\,[(2 - ha_1)(2 + ha_1)^{-1}, \ldots, (2 - ha_N)(2 + ha_N)^{-1}]$($T$ and the a_i are defined in Section I). In view of the relation between the local-truncation errors of the trapezoidal rule and formula (12), this suggests that for nonlinear problems the trapezoidal rule should permit larger step sizes for the same accuracy *when* the "fast components" of the solution have decayed to a very low level.

for example)[4] much work directed toward the comparison of available methods is needed.

REFERENCES

1. Hamming, R. W., *Numerical Methods for Scientists and Engineers,* McGraw-Hill Book Co., New York (1962).
2. Ralston, A., *A First Course in Numerical Analysis,* McGraw-Hill Book Co., New York (1965).
3. Koehler, D., "The Charge-Control Concept in the Form of Equivalent Circuits Representing a Link Between Classic Large Signal Diode and Transistor Models," B.S.T.J., *46,* No. 3 (March 1967), pp. 523–576.
4. Pope, D. A., "An Exponential Method of Numerical Integration of Ordinary Differential Equations," Commun. ACM, *6* (August 1963), pp. 491–493.

Additional References

Branin, F. H., "Computer Methods of Network Analysis," Proc. IEEE, *55,* No. 11 (November 1967), pp. 1787–1801.

Certaine, J., "The Solution of Ordinary Differential Equations with Large Time Constants," chapter 9 of *Mathematical Methods for Digital Computers,* ed. A. Ralston and H. S. Wilf, New York: John Wiley and Sons, 1960.

Dahlquist, C. G., "A Special Stability Problem for Linear Multistep Methods," BIT, *3,* No. 1 (1963), pp. 27–43.

Rosenbrock, H. H., "Some General Implicit Processes for the Numerical Solution of Differential Equations," Computer J., *5,* (January 1963), pp. 329–330.